Computer Speech & Language 64 (2020) 101100

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Vocal tract shaping of emotional speech

Jangwon Kim, Asterios Toutios*, Sungbok Lee, Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Article History:
Received 25 February 2019
Revised 18 December 2019
Accepted 14 March 2020
Available online 16 April 2020

PACS:
43.70.Jt
43.70.Aj
43.72.Ar

Keywords:
Emotional speech production
USC-EMO-MRI Corpus
MR Image segmentation
Vocal tract shaping

ABSTRACT

Emotional speech production has been previously studied using fleshpoint tracking data in speaker-specific experiment setups. The present study introduces a real-time magnetic resonance imaging database of emotional speech production from 10 speakers and presents articulatory analysis results of speech emotional expression using the database. Midsagittal vocal tract parameters (midsagittal distances and the vocal tract length) were parameterized based on a two-dimensional grid-line system, using image segmentation software. The principal feature analysis technique was applied to the grid-line system in order to find the major movement locations. Results reveal both speaker-dependent and speaker-independent variation patterns. For example, sad speech, a low arousal emotion, tends to show smaller opening for low vowels in the front cavity than the high arousal emotions more consistently than the other regions of the vocal tract. Happiness shows significantly shorter vocal tract length than anger and sadness in most speakers. Further details of speaker-dependent and speaker-independent speech articulation variation in emotional expression and their implications are described.

1. Introduction

It has been previously suggested that the emotional state of a speaker influences the neuro-muscular control of vocal organs during speech production, resulting in emotion-dependent acoustic variations in the speech signal. Acoustic characteristics, in particular prosodic and spectral aspects of emotional speech have been extensively studied (Paeschke et al., 1999; Scherer, 2003; Mozziconacci and Hermes, 2000; Banse and Scherer, 1996), but only a few works have addressed articulatory characteristics of emotional speech using vocal tract movement data. Using electromagnetic articulography (EMA), Erickson et al. (1998, 2004, 2006, 2016) found that emotional expression can be characterized by different positions of articulators, such as the tongue tip, the tongue dorsum, the jaw and the lips. For example, they found more peripheral tongue positioning for vowel /i/ in emotional speech than neural speech. Lee et al. (2005) found that emotional speech articulation exhibits more peripheral or advanced tongue positions than neutral speech articulation, and that the movement range of the jaw is larger for angry speech compared to neutral, sad or happy speech. Recently, Ren et al. (2018) reported greater vocal tract opening of happy speech and angry speech than neural speech for Mandarine Chinese words; these patterns are similar to those in American English speech.

The interplay between articulatory movements and voice source activity in emotional speech has also been studied. Kim et al. (2010) found that the difference between happy and angry speech can be better captured by the joint analysis of F0 and articulatory kinematics. Specifically, articulatory variations are more emphasized in angry speech, while F0 variations are more significant in happy speech. This pattern was observed in the horizontal and vertical movements of the tongue tip, the lower lip and the jaw, which were measured using EMA.

*Corresponding author.
E-mail address: toutios@usc.edu (A. Toutios).

https://doi.org/10.1016/j.csl.2020.101100
0885-2308/

Recently, Kim et al. (2015) reported empirical evidence about the relationship between the variability of an articulator and the linguistic criticality of the articulator in emotional speech. An articulator was considered to be linguistically critical if a specific position or the gestural movement of the articulator is crucial for producing a speech sound. For example, the tongue tip is a critical articulator for the alveolar stop /t/, because the closure and releasing movement of the tongue tip is essential for producing the /t/ sound. They reported that linguistically less critical articulators show more emotion-dependent variability than linguistically more critical articulators.

Although all these studies suggest that articulatory movements are affected by the emotional state of the speaker, a possible limitation is that they observed only the movements of a few flesh points inside the vocal tract. This limitation may be addressed by use of real-time Magnetic Resonance Imaging (MRI). Real-time MRI is a non-invasive articulatory data acquisition method that offers a comprehensive view of the dynamics of vocal tract shaping along a plane, typically the mid-sagittal plane (Narayanan et al., 2004). Using real-time MRI data, Lee et al. (2006) reported some preliminary findings on emotion-dependent vocal tract shaping. For example, the vocal tract length and pharyngeal constriction, which cannot be assessed directly using EMA, varied depending on emotions. Their findings were, however, obtained from a limited amount of data from a single male speaker.

The present study investigates articulatory variability in real-time MRI data of ten speakers from a recently collected corpus, that we call the USC-EMO-MRI[1] database. This is a multimodal database of emotional speech, comprising Magnetic Resonance (MR) video data (sequences of upper airway images with synchronized speech audio after noise reduction) and perceptual evaluation results of speech emotional quality. This corpus was designed to serve as a resource in the context of diverse speech production studies, addressing, for example inter- and intra-speaker variability of vocal tract shaping, resultant acoustic variations, and computational modeling of emotional speech production. Kim et al. (2014b) provided a brief summary of the corpus and reported preliminary analysis of articulatory variation across emotions, using a part of a single speaker's data from the database. As an extended-version of the preliminary report, the goal of the present study is three-fold: (i) to provide details about the USC-EMO-MRI corpus, (ii) to re-visit the preliminary findings of articulatory variability in the literature, using data from multiple speakers, and (iii) to discover novel emotion-dependent variability in vocal tract shaping. Even if they cannot be generalizable to the overall population, our findings suggest consistency of several previous works on the topic, and report new articulatory characteristics of emotional speech. The USC-EMO-MRI corpus is publicly and freely available for research purposes at http://sail.usc.edu/span/usc-emo-mri with real-time MRI video (MR image sequences and speech audio) and perceptual emotion evaluation results for each utterance.

The present study investigates vocal tract shaping of emotional speech in terms of a distance function and the vocal tract length. The distance function refers to the collection of the Euclidean distances between inner and outer tissue-airway boundaries in the oropharyngeal vocal tract as a function of the distance from the lips. The distance function has been popularly used as a set of initial parameters for vocal tract shaping in the literature (Öhman, 1967; Story, 2009; Proctor et al., 2010). The vocal tract length refers to a curvilinear line equidistant from the inner and outer tissue-airway boundaries in the oropharyngeal vocal tract. In the present study, we computed these vocal-tract-shape parameters automatically, using a segmentation algorithm that we have developed (Kim et al., 2014a). This algorithm performs tracking of the lips and the larynx, and detection of the tissue-airway boundary points in vocal-tract grid lines that are systematically spread over the oropharyngeal vocal-tract space. The details of this algorithm are provided in Section 3.1(for the distance function) and Section 3.3 (for the vocal tract length).

The distance function is generally redundant (its elements are highly correlated) due to the physiological constraints of the vocal tract (e.g., smooth shape of the surface) and the coordinated controls of speech articulators. Hence, previous studies often reduce the number of parameters (in the distance function) as pre-processing for their analysis and modeling of the vocal tract, by using decomposition techniques (Liljencrants, 1971; Harshman et al., 1977; Story et al., 1996; Yehia et al., 1996; Story and Titze, 1998; Mokhtari et al., 2007; Cai et al., 2009). The conventional decomposition methods, e.g., Principal Component Analysis (PCA) and Fourier series, transform the initial parameters to a low-dimensional, compact set of parameters. However, the behavior of each parameter in the low-dimensional space is difficult to interpret, because the low-dimensional space is hidden. Even if we project the behavior of each low-dimensional parameter to the original parameter space, the influence of the variation in a hidden parameter to (interpretable) initial parameters is complex. These issues make the analysis of 'local variability' of vocal tract shaping challenging.

We employ Principal Feature Analysis, or PFA (Lu et al., 2007) for a compact representation of the distance function. PFA selects the most variable and least redundant locations in the vocal tract directly, which allows for an easily interpretable and compact feature set of vocal tract shaping. Analyzing the most variable vocal-tract locations is important for the present study, because variations of articulatory movement range are important emotion-dependent behaviors (Lee et al., 2005; Kim et al., 2010; 2015). The present study examines the local variability of vocal tract shaping in emotional speech, which can be easily captured by the PFA method. The details of PFA are provided in Section 3.2.

Using the PFA method, we examine our hypothesis that emotion-dependent variations relative to neutrality can be captured in the principal feature space. We also examine what emotion-dependent variations are speaker-dependent and speaker-independent in the USC-EMO-MRI corpus.

[1] The USC-EMO-MRI is an acronym for University of Southern California (the organization of authors) - Emotion (interested quality of speech) - Magnetic Resonance Imaging (articulatory data collection modality).

This paper is organized as follows: Section 2 describes the USC-EMO-MRI corpus. Section 3 describes how we computed vocal tract parameters from the real-time MRI data. Section 4 reports analysis results for the emotion-dependent variations of vocal tract shaping in terms of the principal features and the vocal tract length. Section 5 provides discussion.

2. The USC-EMO-MRI corpus

The USC-EMO-MRI corpus comprises real-time MRI data and corresponding speech audio from five female and five male speakers, and categorical emotion labels for each utterance. All speakers have had earlier professional acting experience and theatrical vocal training. It is noted that we analyzed vocal tract shaping using a subset (one sentence) of this dataset as an illustrative investigation.

2.1. Speech stimuli

Table 1 shows the list of speech stimuli which speakers read during data recording, with each target emotion. The target emotions consist of three basic emotions of happiness, sadness and anger, and a neutral emotion. The set of sentences was designed to investigate the effects of emotional expression on prosodic and rhythmic structure in articulatory movements, including reiterant speech (Kelso et al., 1985). Reiterant speech (Sentence 7 in Table 1) is the speech in which each various syllables in each word in the original sentence (Sentence 6 in Table 1) is replaced by a consonant-vowel syllable, such as "Ma," and the original intonation is being maintained.

Data collection from each speaker consisted of four parts. The speakers were asked to immerse themselves in a different emotion throughout each part. Speakers read the passage in normal speaking rate for all four emotions (including the neutral emotion) and additionally in fast rate only for the neutral emotion. They also read six sentences in seven repetitions for each of the four emotions, all in normal speaking rate. The order of presentation of the sentences was randomized in each repetition. For some speakers, a seventh "nonsense" sentence (reiterant speech) was added. That sentence was always presented right after Sentence 6 in Table 1 and speakers were asked to read it with the same intonation with Sentence 6. When reading Sentence 4, speakers were asked to emphasize the uppercased words in the stimuli, viz. "KNIGHT" and "MONSTER."

Data of Sentence 6 in Table 1 was used for analysis in this study.

2.2. Data acquisition and processing

We used the MRI data acquisition and processing protocols of the USC-TIMIT database (Narayanan et al., 2014). This section offers a summary of the acquisition and processing protocols. See Narayanan et al. (2014) for technical details.

We collected upper airway MR images at the Los Angeles County Hospital using a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha WI). We used a custom 4-channel receiver coil array, with two anterior coil elements and two coil elements posterior to the head and the neck. We recorded the MRI data of speakers while they lay supine in the scanner and read the stimuli.

The real-time MRI acquisition was performed using a spiral fast gradient echo sequence, where thirteen interleaved spirals form a single image. We used a sliding window technique (Narayanan et al., 2004) which allows view sharing, thus increases frame rate. The repetition time, or TR at data acquisition was 6.164 for each spiral, and the TR-increment for view sharing was 7 acquisitions (Narayanan et al., 2004; Bresch et al., 2008; Kim et al., 2011b). Hence, MRI movies were generated with a frame rate of 23.18 frames/sec (1 / (7 $\times$ 6.164 msec)). The field of view of imaging was 200 $\times$ 200 mm, and image resolution was 68 $\times$ 68 pixels (2.94 $\times$ 2.94 mm for each pixel). We used RTHawk (HeartVista, Inc., Los Altos, CA), which is a custom real-time imaging platform (Santos et al., 2004), for the scan plane localization of the mid-sagittal slice. We recorded speech audio at a sampling frequency of 20kHz, simultaneously with MR imaging, using a custom fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor,

Table 1
List of speech stimuli for the USC-EMO-MRI corpus.

Sentence 1	John bought five black cats at the store.
Sentence 2	The leopard, skunk and peacock are wild animals.
Sentence 3	Charlie, did you think to measure the tree?
Sentence 4	The queen said the KNIGHT is a MONSTER.
Sentence 5	Hickory dickory dock, the mouse ran up the clock. Hickory dickory dock.
Sentence 6	9 1 5 (short pause) 2 6 9 (short pause) 5 1 6 2.
Sentence 7	Ma Ma Ma (short pause) Ma Ma Ma (short pause) Ma Ma Ma Ma.
Grandfather passage	You wished to know all about my grandfather. Well, he is nearly ninety-three years old; he dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

Israel) and a custom recording setup; the unblank Transistor-Transistor Logic (TTL) signal from the scanner, which is generated at the beginning of each MRI acquisition, triggers audio data recording.

During post-processing, we performed noise cancellation on speech audio using a custom adaptive signal processing algorithm (Bresch et al., 2006), followed by the synchronization of the MRI video and speech audio.

2.3. Evaluation of emotional quality

Perceptual evaluation tests were performed to assess the emotional quality of the recorded data, i.e., how well the intended emotion (by speakers) is expressed in speech audio. Ten to twelve evaluators tested each speaker's data via headphones. (See Table 2 for exact numbers of evaluators.) They were allowed to listen to as many times as they wanted. Some of the evaluators were also actors or actresses who participated in the data collection. After listening to the recorded audio for each utterance, the evaluators were asked to: (i) choose the emotion that they perceive from the spoken sentence (neutral, anger, happiness, sadness, or other); (ii) judge their confidence in their choice; and (iii) judge the strength of the emotional expression. Confidence and strength were evaluated on a five-point Likert scale. Only two speakers (M2 and F2) out of ten speakers performed self-evaluation on their speech data. The difference between with self-evaluated emotion labels and without self-evaluated emotion labels was not significant in this dataset, because the number of self-evaluation is too small (only one of 10 evaluation results for M2 and one of 12 for F2).

Table 2 also presents the number of sentences recorded from each speaker and the average and standard deviation of the matching ratio between target emotion (intended emotion by the speaker) and the final emotion label (perceived emotion by listeners). The final emotion label was determined by majority voting across all evaluators. The matching ratio counts the number of the utterances whose target emotion and the final emotion label match over the number of all utterances.

The average agreement rates vary between 94.0% (F2) and 69.5% (M2) in Table 2. This is possibly due to the mixed effect of (i) differences in the goodness of target emotion portrayals among speakers and (ii) differences in perception among listeners. Also, the agreement rates of female speakers' data are greater than those of male speakers' data, which implies that the female speakers may have portrayed target emotions better than the male speakers.

3. Methods

3.1. MR Image parameterization

A tool for the automatic tissue segmentation in real-time MR images has been developed for the purposes of this study. Kim et al. (2014a) presented an earlier version of the implemented method which shows more accurate segmentation performance across oropharyngeal airway compared to previous work (Proctor et al., 2010). This method performs image enhancement first in a retrospective manner as follows: Local pixel intensity map was computed from reconstructed MR image sequences first, then the value of the map was used to normalize pixel intensity of MR images. Then, it seeks pixel intensity thresholds distributed along tract-normal grid-lines and defines airway contours constrained with respect to an estimated airway path from the glottis to the lips.

Fig. 1 illustrates the outputs of MR image segmentation procedures, using an MR image of speaker M1. The method is initialized with a manually drawn reference line and manually set points. The reference line serves as a reference for determining the center line of grid lines. Our tool provides a standard deviation image (image in Fig. 1(a)), so that the user can draw the line easily on the image by roughly following the outer boundary of highlighted pixels in the oropharyngeal airway. Generally, the resulting line is optimally placed near the upper tissue-airway boundary in the front oral cavity and the back pharyngeal wall in the back cavity, so that it is part of the airway in most MR image frames. Finally, the reference line (yellow dots in Fig. 1(a)) is determined by smoothing the line, using the discrete cosine transform. The manually set points (red dots in Fig. 1(b)) correspond to the midpoint of the lips, the highest point on the palatal surface, and the upper visible boundary of the arytenoid cartilage. The manual points for the lips and the arytenoid cartilage are used as reference points to determine the regions of search for the initial and

Table 2
Summary of evaluation results of all evaluators. 'Sentences' indicates the sentence ID included. 'Average' and 'STD' denotes average and standard deviation of the matching ratio (%) between target emotion and the final emotion label for sentence-level utterances, respectively. This table is adopted from (Kim et al., 2014b).

	Subject ID (M: male, F: female)									
	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
# Evaluators	10	10	11	10	11	12	12	12	12	10
Sentences	1-6	1-7	1-7	1-7	1-7	1-6	1-6	1-6	1-7	1-7
Average	85.3	69.5	82.6	72.0	80.5	80.7	94.0	89.8	86.5	80.5
STD	9.9	11.8	8.5	11.1	10.0	11.1	5.4	8.4	8.2	11.8

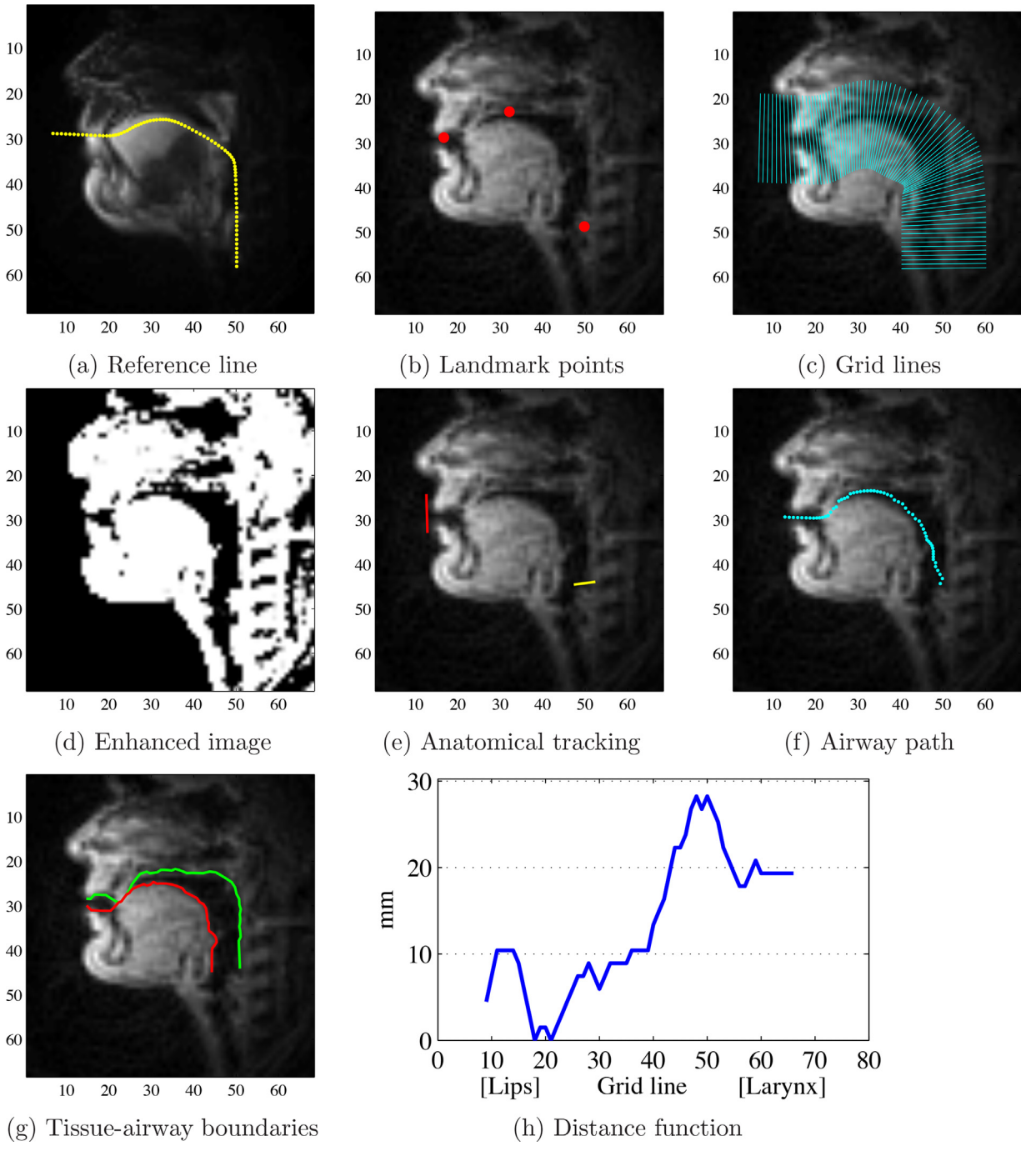

Fig. 1. Results of parameterization processes of a magnetic resonance image (speaker M1) as an example. X-y axes in (a)-(g) indicate the pixel index.

the final boundaries in the vocal tract, respectively. The manual point on the palatal surface is used for determining a region of search for the oropharyngeal airway (the algorithm searches below the manually selected palatal point). The tool uses the arytenoid cartilage instead of the glottis, because the glottis is not well imaged in the MR images. This initialization is done once; the upper boundary of the arytenoid cartilage in each frame is detected automatically, and so is the exit of the lips.

A set of equidistance grid-lines is constructed automatically, perpendicular to the reference line. This is shown in Fig. 1(c). After retrospective MR image quality enhancement (Fig. 1(d)), the first of these grid-lines is located at the (automatically detected) upper boundary of the arytenoid cartilage and the last at the exit of the lips, as shown in Figure (e). A frame-specific

airway path from the first to the last grid-line is then determined, using dynamic programming algorithm, followed by spatio-temporal smoothing. The final airway path is shown in Fig. 1(f). It is noted that the airway path can be any line passing through the oropharyngeal airway in the present paper. Then, for each grid-line inner and outer tissue-airway boundaries are determined; The inner boundary is a line on the surface of the lower lip and tongue, while the outer boundary is a line on the surface of the upper lip, the alveolar ridge, the palate and the pharyngeal wall. Initial inner and outer boundary points for each grid line are determined as the first pixels whose intensity is above a certain threshold (0.5), along the grid-line and in toward the two grid-line edges starting from the intersection of the airway path and the grid-line. The boundaries are shown in Fig. 1(g).

The method was applied on all non-silent regions of the previously described corpus. Silent regions were determined using SailAlign (Katsamanis et al., 2011) which is a Hidden Markov Model based adaptive speech-to-text forced aligner. The method was initialized whenever significant differences in head posture (high, low, forward, backward), head tilting, or neck stretching, were observed. The Euclidean distances between the two airway-tissue boundaries on each grid-line were measured. Fig. 1(h) shows the Euclidean distance for each grid line in the vocal tract for the example MR image in the figure. The final vocal tract parameters computed from the distances (i.e., distance function), using PFA, is discussed in the following section.

3.2. Principal feature analysis

For a compact representation of vocal tract shaping, we performed feature reduction using PFA (Lu et al., 2007). This method selects a subset of the original features (i.e., distances between outer and inner boundaries on the grid lines), using the same feature reduction criteria as PCA. Sample points are maximally spread in the selected features, where the structure of the principal components is retained, thus preserving the variation of the original data. Hence, PFA offers a compact, effective and interpretable representation of vocal tract shaping.

We computed the principal features of individual speaker's data as follows: (1) Compute eigenvectors and eigenvalues from the covariance matrix of distance functions. (2) Choose the minimum number of dimension, q for the subspace, where the cumulative sum of eigenvalues for the dimensions is greater than 90% of the total sum. (3) Perform k-means clustering on the row vectors of the subspace eigen matrix. In our case, k was $3-7$ greater than q for retaining 90% of the total variation in the subspace. (4) Find the row vector corresponding to the mean of each cluster, where the index of the row vector becomes the index of the selected original feature. Steps (1) and (2) are identical to PCA computation, which are for efficiently preserving the variation of the original data. Steps (3) and (4) are extra procedures for PFA, compared to PCA, which are for selecting less dependent ones among the original features. See Lu et al. (2007) for full details of this algorithm.

Fig. 2 shows the subset (principal features) of the grid lines overlaid on an arbitrarily chosen midsagittal image of the corresponding speaker. There are more principal features in the front oral cavity (lips to the hard palate) than pharyngeal region, implying more complex movements of the tongue in the front oral cavity than the tongue in the pharyngeal region. Some principal features are sometimes clustered closely, especially near the alveolar ridge and the teeth. It is hypothesized that the delicate movements of the tongue tip cause neighboring grid lines in the region to be less correlated.

Next, we performed temporal alignment of utterances in order to compare time series of principal features for different emotions. Each utterance was mapped to a reference utterance, using a dynamic time warping algorithm (Sakoe and Chiba, 1978) The reference utterance was arbitrarily selected from data of neutral emotion. Fig. 3 illustrates the time series of principal features before and after the temporal alignment using data of individual speakers (M1 in this figure). Finally, we computed averaged time series for each emotion, which were used to specify the analysis of emotion-dependent variation in vocal tract shaping along the midline of the vocal tract.

3.3. Vocal tract length computation

The computation of the true vocal tract length requires the location of the lips (the initial point of the vocal tract), the location of the true vocal fold (the final point of the vocal tract), and the center line between outer and inner tissue-airway boundaries in the vocal tract. We selected the grid line of the smallest distance in the lip region for the initial point of the vocal tract. We selected the grid line of the top of the arytenoid cartilage for the final point of the vocal tract, because, as mentioned in Section 3.1, the true vocal fold (glottis) is not well imaged in our dataset. We computed the geodesic distance (the sum of the Euclidean distance between the center points of adjacent grid lines) within outer and inner boundaries from the initial point to the final point, which is considered as an approximation of the vocal tract length in the context of this study. Next, we excluded samples whose values were out of empirically selected upper and lower boundaries for each speaker. We set the upper boundary to be mean + 3 $\times$ standard deviation, and the lower boundary to be mean - 3 $\times$ standard deviation. Then, we computed the statistics of the vocal tract length in order to compare how the vocal tract length varies depending on emotion. It should be noted that we aligned the time-series of the vocal tract length before computing the statistics. This process minimizes vocal tract length variations due to different durations of phones in speech production in different utterances. The outputs of the alignment in Section 3.2 were used here as well.

The MATLAB tool that computes distance function, principal features, and the vocal tract length from MR images is freely available at http://sail.usc.edu/old/software/rtmri_seg

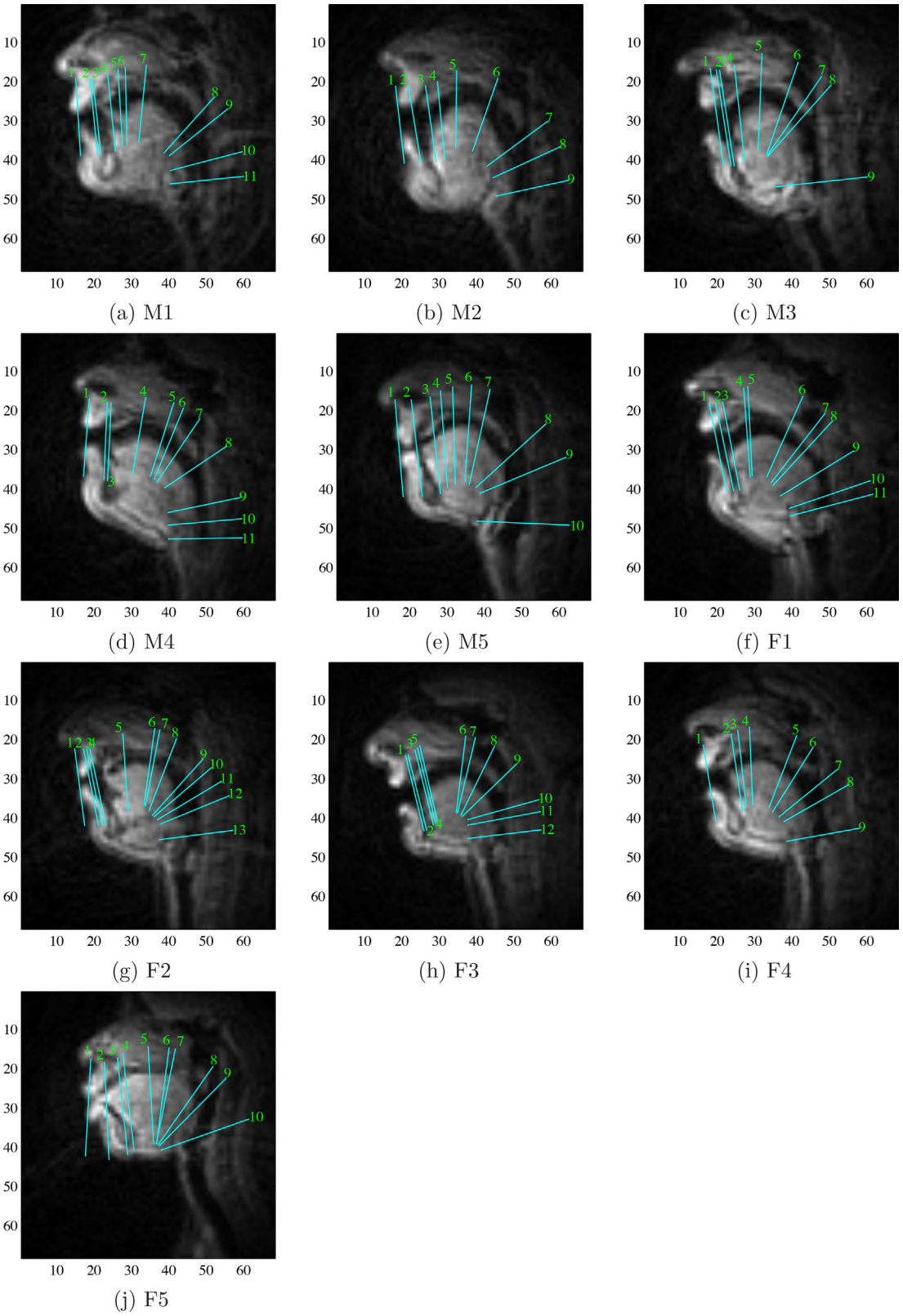

Fig. 2. Principal features overlaid on an MR image for each speaker. Cyan lines are the principal features. Green numbers are the principal feature indices.

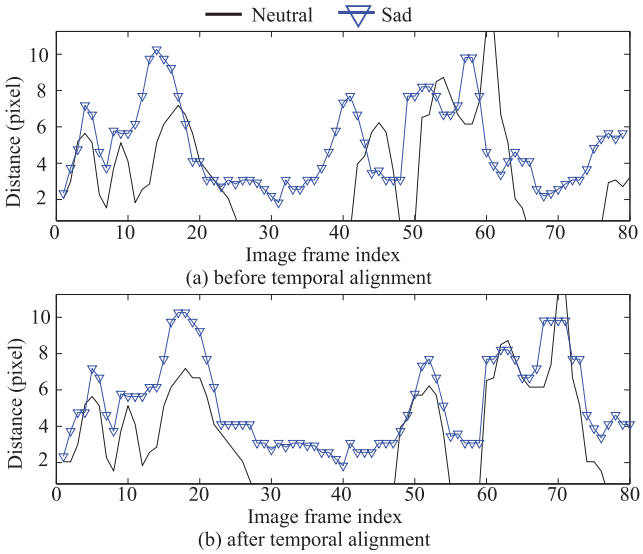

Fig. 3. Time series of the third principal feature before (a) and after (b) temporal alignment of an utterance to a reference (neutral) utterance of speaker M1.

4. Results

4.1. Emotion-dependent variations in principal features

This section discusses emotion-dependent variations captured in principal features. From this point on, we report within-speaker analyses across emotions.

As an initial investigation, we compared the averaged time series of principal features for each of the four emotions. Fig. 4 illustrates the averaged time series of two principal features for speaker M1 as examples. Overall, the time series clearly show differences depending on emotions. For the first principal feature (located at the lips) in Fig. 4(a), the distances of anger and happiness are often greater than those of neutrality and sadness. It is noted that anger and happiness are high arousal emotions, while sadness is a low arousal emotion (Kim et al., 2014b). Hence, this indicates that speaker M1 shows positive correlation between the degree of his lip opening and the arousal dimension of emotion. For the seventh principal feature (located on the hard palate) in Fig. 4(b), the distances for sadness are mostly greater than those for the other emotions, especially from the 20th frame to the 35th image frame. This region corresponds to the words “two six” that include high vowels /u/ and /I/. This suggests that the vertical constriction gesture of the tongue dorsum tends to be less strictly controlled for sadness, compared to the other emotions,

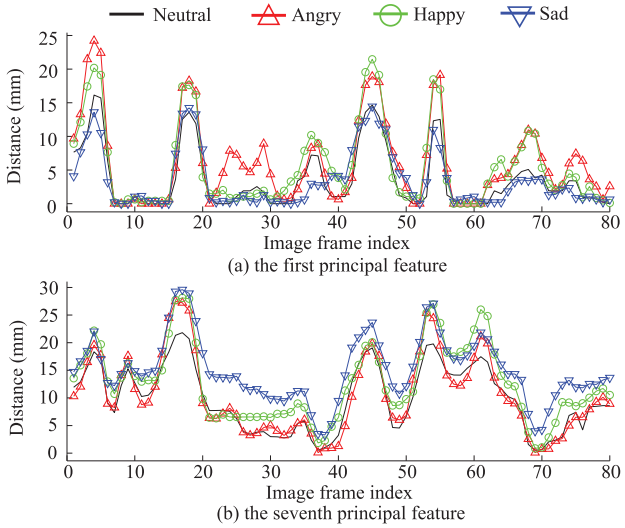

Fig. 4. Averaged time series of the first and the seventh principal features for each emotion of speaker M1. The averaged time series were temporally aligned. The utterances of Sentence 6 “nine one five two six nine five one six two” are used.

when the speaker utters the high vowels. Interestingly, happiness and anger show clearly different patterns in the range of the first principal feature for /u/, that is located near the image frames 23 and 75, while they show similar patterns in the seventh principal feature.

Based on the observation of the emotion-dependent patterns in the time series of principal features, we analyzed emotion-dependent variations relative to neutrality for all principal features for each speaker. First, we computed statistics ([0.1, 0.5, 0.9] quantiles, and 0.9 quantile - 0.1 quantile) of each principal feature and each emotion. 0.1 quantile, 0.5 quantile and 0.9 quantile reflect constriction, median position and large opening in the local vocal tract area of the corresponding principal feature, respectively. For example, constriction gestures for high vowels, labial and coronal consonants are reflected by 0.1 quantile of the principal features in the palatal and lip regions. Large opening gestures for low vowels are reflected by 0.9 quantile of the principal features in the palatal region. 0.9 quantile - 0.1 quantile of a principal feature reflects movement range in the corresponding vocal tract area. Next, we computed relative statistics by subtracting the statistics of each emotion from those of a neutral emotion so that the relative statistics capture how each emotion perturbs vocal tract shaping. It is noted that we computed [0.1, 0.5, 0.9] quantiles instead of minimum and maximum in order to minimize the effects of (possible) tissue-airway segmentation errors on this analysis.

From here, we will discuss both speaker-dependent and speaker-independent variation patterns in vocal tract shaping, captured in principal features. It is noted that the location of each principal feature in the vocal tract is shown in Fig. 2.

Fig. 5 shows results of 0.1 quantile (constriction) for each speaker. First, smaller (narrower) or similar constriction in the vocal tract for anger than sadness are observed for nine speakers (except speaker F5 in Fig. 5(j)), although the regions of smaller constriction vary depending on speakers. For example, speaker M1 (Fig. 5(a)) and speaker M2 (Fig. 5(b)) show smaller constrictions in principal features 5 - 7 and principal features 2, 4 and 5, respectively, where these principal features are located in the front oral cavity for both speakers. However, speaker F4 (Fig. 5(i)) shows smaller constriction in principal features 7 and 8, which are located in velar and pharyngeal regions. Second, wider or similar constrictions are observed for sadness than neutral in most of the vocal tract regions for most speakers (M1, M3, M4, M5, F2, F3 and F4). However, rather narrower constrictions are observed for sadness than neutral in the pharyngeal regions or near the larynx for the other speakers (M2, F1 and F5). Third, although clearly narrower constriction for anger than neutral is observed in principal features 2, 4 and 5 for M2, many speakers show both smaller and greater (wider) constrictions. For example, M4 shows smaller constriction for anger than neutral in principal features 5 and 6, while greater constriction in principal feature 9.

Fig. 6 shows results of 0.9 quantile, reflecting large opening in the specific vocal tract regions. First, larger or similar openings are observed in the most principal features for high arousal emotions (anger and happiness) than neutral. This indicates that the ten speakers in the USC-EMO-MRI corpus tend to emphasize opening gestures for high arousal emotions. However, compared to neutral, sadness show different (larger, similar and/or smaller) openings depending on speakers and principal features. For example, F5 shows mostly smaller openings for sadness than neutral in all principal features, while F4 shows greater openings in the principal features 5, 6, 7 and 8 (the velar-pharyngeal region). This indicates that the speakers have different levels of emphasis on opening gestures for sadness. Finally, anger and happiness mostly show larger or similar openings in the front oral cavity than sadness. However, this pattern is not consistent in the velar and pharyngeal regions; sadness shows greater openings than happiness significantly in principal features 5, 6, 7 and 8 for F4, and slightly in the principal features 7, 8, 10, 11 for F3. Similar patterns to Fig. 6 were observed in the results for 0.9 quantile - 0.1 quantile (not shown in the present paper). Also, results of 0.5 quantile showed some emotion-distinctive patterns, but speaker-independent patterns were not found.

4.2. Emotion-dependent variations of the vocal tract length

Fig. 7 shows boxplots of the vocal tract length for each emotion and each speaker. It is observed that the vocal tract length also varies depending on emotions. Specifically, the vocal tract length tends to be shorter for happiness than anger or sadness across the ten speakers. We conducted one-tailed Welch's t-test on the hypothesis that the mean of the vocal tract length for happy speech is shorter than the mean for angry/sad speech. Results indicate that on average, happy speech shows statistically significantly shorter vocal tract length than angry speech at $\alpha = 5 \times 10^{-6}$ level in general, except for M5 (t-statistic = 1.64, $p = 0.05$). Also, happy speech shows statistically significantly shorter vocal tract length than sad speech for all speakers' data ($p < 0.00$).

5. Discussion

Based on the USC-EMO-MRI database and the MR image tracking software, the present study examined how the vocal tract shape changes depending on emotions, using automatically extracted vocal tract parameters which describe the distance between the inner and outer vocal-tract boundaries and (approximate) vocal tract length.

First, the present study provided supporting evidence for previous findings by Lee et al. (2006) for emotion-dependent variation patterns, using lexically richer data from the ten speakers in the USC-EMO-MRI corpus. Lee et al. (2006) found that anger shows wider opening in both oral cavity and pharyngeal region than neutrality. The result of the wider opening in the oral cavity for anger than neutrality was consistent with (Scherer, 1986; Murray and Arnott, 1993; Lee et al., 2005; Kim et al., 2010; 2011a), where the movement range of a sensor attached on the tongue tip was analyzed. The present study found that high arousal emotions (both happiness and anger) show greater movement range than neutrality, but the vocal tract region of significantly contrasted opening varies depending on speakers. Smaller opening of sad speech is also in the line with previous findings by Kienast et al. (1999), Erickson et al. (2004, 2006).

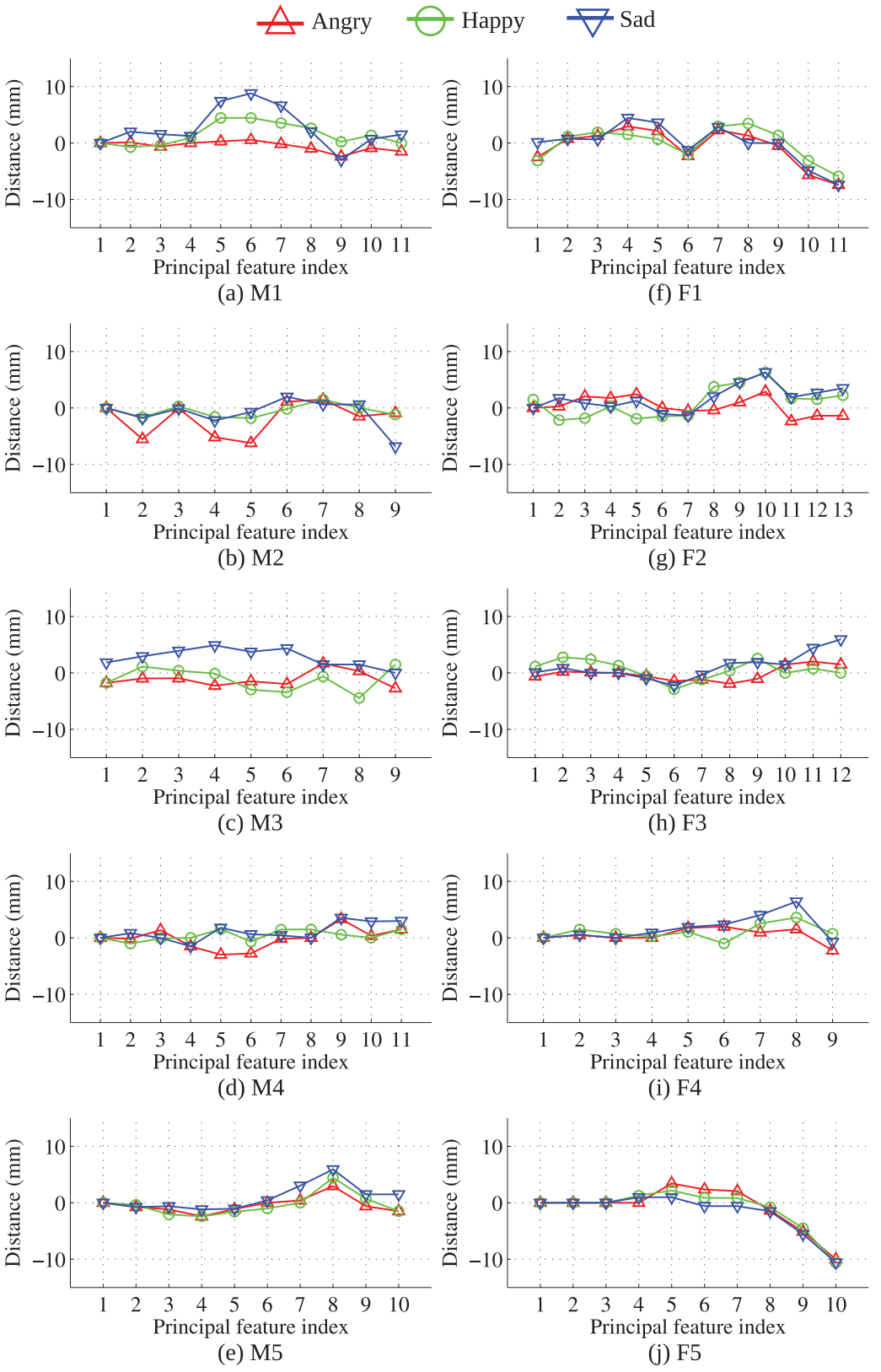

Fig. 5. 0.1 quantiles of principal features for anger, happiness and sadness relative to neutrality for each speaker.

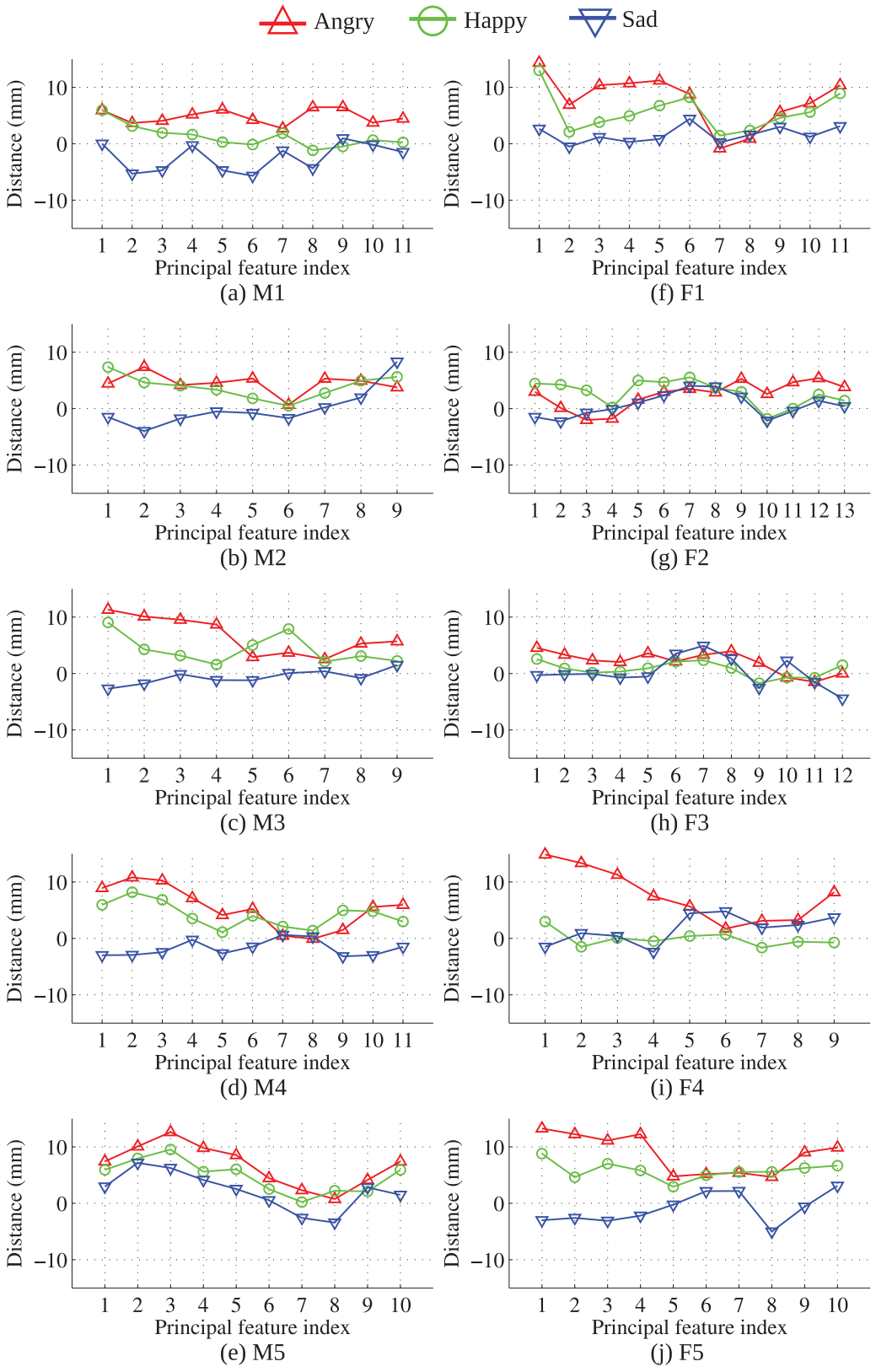

Fig. 6. 0.9 quantiles of principal features for anger, happiness and sadness relative to neutrality for each speaker.

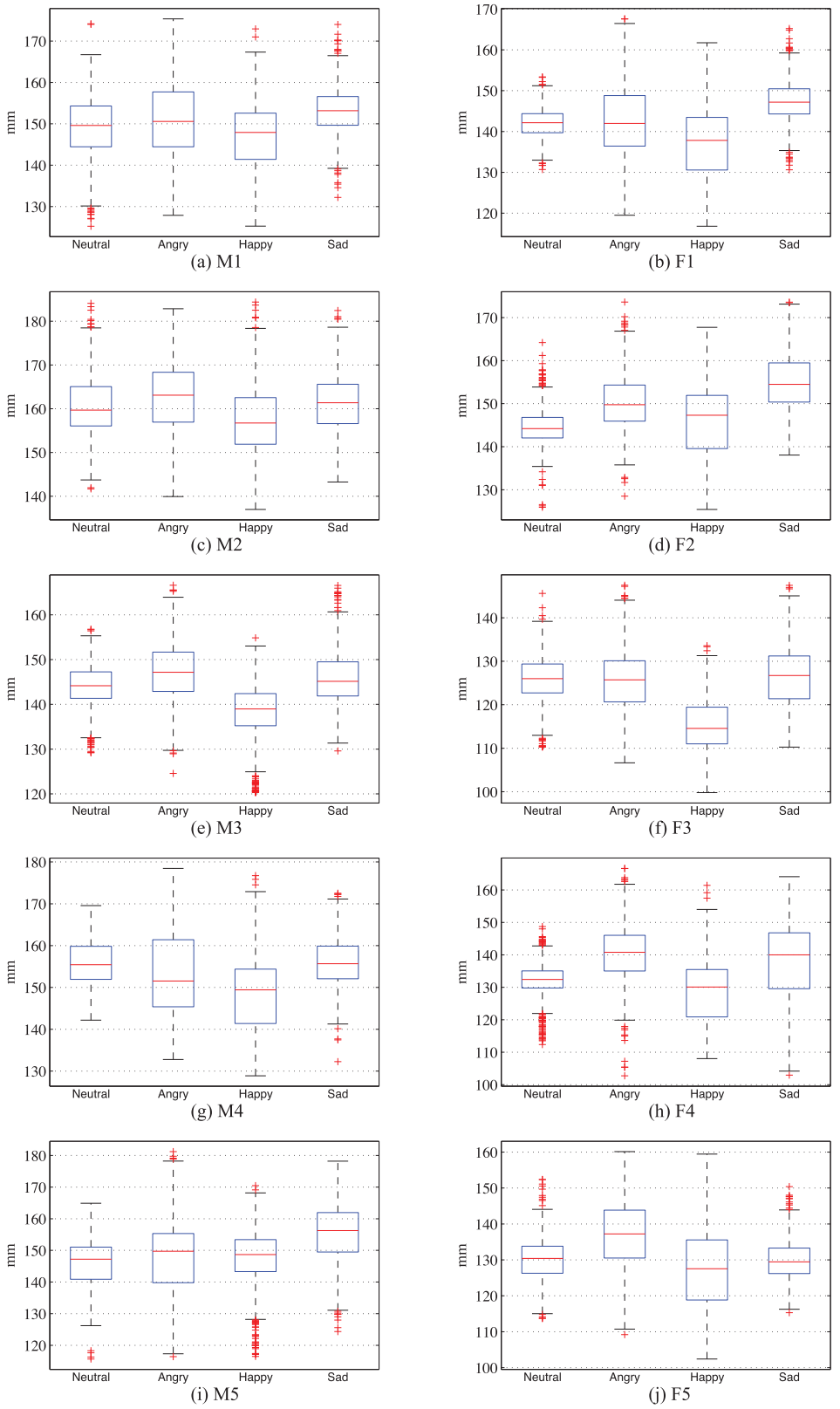

Fig. 7. Boxplots of the vocal tract length of each emotion.

Also, the pattern that happy speech shows shorter vocal tract length than angry, neutral and sad speech is consistent with (Lee et al., 2005). Simulation experiments by Xu and Chuenwattanapranithi (2007) have suggested that the dynamic variations of the vocal tract length (and F0 jointly) are often perceived as expression of joy or anger, but depending on vowels. While the previous studies focused on specific phones, the present study provides evidence that the relationship between the vocal tract length and emotional quality holds for continuous speech as well; this was consistent across speakers. In addition, lip spreading and larynx elevation are important factors contributing to the decreasing of the vocal tract length for happy speech. Tartter (1980) The relative contributions of these factors have been explored by Lasarcyk and Trouvain (2008), jointly with F0 raising, using isolated synthetic vowel sounds. That study found that lip spreading and laryngeal elevation often affect perceived emotional quality in the dominance dimension (representing apparent strength or power of the speaker over the other). Similar simulation experiments in terms of perception of happiness would also be useful to understand the influence of the individual factors to emotional expression. We do not include simulation experiment in the present study due to the difficulty of robust spectral feature extraction from the speech audio in the USC-EMO-MRI corpus; although speech intelligibility is much enhanced by post-processing described in Section 2.2, it still suffers from residual noise and/or post-processing artifact.

The present study also reports a novel finding that when low vowels are produced, sadness shows the smaller opening in the front oral cavity than anger and happiness. The articulatory characteristics for sadness have been studied by Erickson et al. (2004, 2006), where stronger constriction gesture for a high vowel /i/ was observed in sad speech than neutral speech. However, such stronger lingual gestures for sad speech are not consistent for all speakers in the USC-EMO-MRI corpus. For example, Fig. 5 shows weaker constriction gestures in the oral cavity for sad speech than neutral speech. In fact, stronger lingual gestures for any particular emotion across all speakers were not observed. It could be speaker-specific, but more investigation is needed to understand further any such difference.

The vocal tract length is considered as an important morphological parameter which contains speaker-specific information. For example, Smith and Patterson (2005) found the vocal tract length is an important cue for differentiating speaker size, sex and age. The present study suggests that the vocal tract length is an important cue for differentiating the emotional state of the speaker. In fact, Kockmann et al. (2011) reported that normalizing vocal tract length when computing acoustic features does not improve the accuracy of emotion recognition from the speech signal, although this normalization technique has been generally useful for reducing speaker variability (Lee and Rose, 1996). However, in order to use (predicted) vocal tract length parameter as an emotional cue, we need to understand how to decompose its variation into emotional and other speaker-dependent factors, such as age and gender. This is an open question that we will pursue in future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by NSF IIS-1116076 and NIH R01DC007124. Portions of this work were presented in “USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging” 10th International Seminar on Speech Production (ISSP), Cologne, Germany, 2014.

References

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70 (3), 614–636.
Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., Narayanan, S., 2008. Seeing speech: capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. Signal Process. Mag. IEEE 25 (3), 123–132.
Bresch, E., Nielsen, J., Nayak, K.S., Narayanan, S.S., 2006. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. J. Acoust. Soc. Am. 120 (4), 1791–1794.
Cai, J., Laprie, Y., Busset, J., Hirsch, F., 2009. Articulatory modeling based on semi-polar coordinates and guilded PCA technique. In: Proceedings of Interspeech. ISCA, Brighton,UK, pp. 56–59.
Erickson, D., Fujimura, O., Pardo, B., 1998. Articulatory correlates of prosodic control: emotion and emphasis. Lang. Speech 41 (3–4), 399–417.
Erickson, D., Menezes, C., Fujino, A., 2004. Some articulatory measurements of real sadness. In: Proceedings of Interspeech. ISCA, Korea, pp. 1825–1828.
Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., Shibuya, Y., 2006. Exploratory study of some acoustic and articulatory characteristics of sad speech. Phonetica 63 (1), 1–25.
Erickson, D., Zhu, C., Kawahara, S., Suemitsu, A., 2016. Articulation, acoustics and perception of Mandarin Chinese emotional speech. Open Linguist. 2 (1).
Harshman, R., Ladefoged, P., Goldstein, L., 1977. Factor analysis of tongue shapes. J. Acoust. Soc. Am. 62 (3), 693–707.
Katsamanis, A., Black, M., Georgiou, P.G., Goldstein, L., Narayanan, S.S., 2011. SailAlign: robust long speech-text alignment. Workshop on New Tools and Methods for Very-Large Scale Phonetics Research. Philadelphia, PA.
Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E.L., Kay, B., 1985. A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling. J. Acoust. Soc. Am. 77 (1), 266–280. https://doi.org/10.1121/1.392268.
Kienast, M., Paeschke, A., Sendlmeier, W.F., 1999. Articulatory reduction in emotional speech. In: Proceedings of Eurospeech, pp. 117–120.
Kim, J., Kumar, N., Lee, S., Narayanan, S., 2014. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. the 10th International Seminar on Speech Production (ISSP). Cologne, Germany, pp. 222–225.
Kim, J., Lee, S., Narayanan, S.S., 2010. A study of interplay between articulatory movement and prosodic characteristics in emotional speech production. In: Proceedings of Interspeech. ISCA, pp. 1173–1176.
Kim, J., Lee, S., Narayanan, S.S., 2011. An exploratory study of the relations between perceived emotion strength and articulatory kinematics. In: Proceedings of Interspeech. ISCA, pp. 2961–2964.

Kim, J., Toutios, A., Kim, Y.-C., Zhu, Y., Lee, S., Narayanan, S., 2014. USC-EMO-MRI corpus: an emotional speech production database recorded by real-time magnetic resonance imaging. the 10th International Seminar on Speech Production (ISSP). Cologne, Germany, pp. 226–229.
Kim, J., Toutios, A., Lee, S., Narayanan, S.S., 2015. A kinematic study of critical and non-critical articulators in emotional speech production. J. Acoust. Soc. Am. 137 (3), 1411–1429. https://doi.org/10.1121/1.4908284.
Kim, Y.-C., Narayanan, S.S., Nayak, K.S., 2011. Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order. Magn. Reson. Med. 65 (5), 1365–1371.
Kockmann, M., Burget, L., et al., 2011. Application of speaker-and language identification state-of-the-art techniques for emotion recognition. Speech Commun. 53 (9), 1172–1185.
Lasarcyk, E., Trouvain, J., 2008. Spread lips+raised larynx+higher F0=Smiled Speech? -An articulatory synthesis approach. In: Proceedings of International Symposium of Speech Production, Strasbourg, France, pp. 43–48.
Lee, L., Rose, R.C., 1996. Speaker normalization using efficient frequency warping procedures. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. 353–356.
Lee, S., Bresch, E., Adams, J., Kazemzadeh, A., Narayanan, S.S., 2006. A study of emotional speech articulation using a fast magnetic resonance imaging technique. In: Proceedings of Interspeech. ISCA, Pittsburgh, PA, pp. 2234–2237.
Lee, S., Yildirim, S., Kazemzadeh, A., Narayanan, S.S., 2005. An articulatory study of emotional speech production. In: Proceedings of Interspeech. ISCA, pp. 497–500.
Liljencrants, J., 1971. Fourier series description of the tongue profile. Speech Trans. Lab. 12 (4), 9–18.
Lu, Y., Cohen, I., Zhou, X.S., Tian, Q., 2007. Feature selection using principal feature analysis. In: Proceedings of the 15th International Conference on Multimedia. ACM, New York, NY, USA, pp. 301–304.
Mokhtari, P., Kitamura, T., Takemoto, H., Honda, K., 2007. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients. J. Phon. 35 (1), 20–39. https://doi.org/10.1016/j.wocn.2006.01.001.
Mozziconacci, S.J., Hermes, D.J., 2000. Expression of emotion and attitude through temporal speech variations. In: Proceedings of Interspeech. ISCA, pp. 373–378.
Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Am. 93 (2), 1097–1108.
Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D., 2004. An approach to real-time magnetic resonance imaging for speech production. J. Acoust. Soc. Am. 115 (4), 1771–1776.
Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., Proctor, M., 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). J. Acoust. Soc. Am. 136 (3), 1307–1311. https://doi.org/10.1121/1.4890284.
Öhman, S.E.G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Am. 41 (2), 310–320. https://doi.org/10.1121/1.1910340.
Paeschke, A., Kienast, M., Sendlmeier, W., 1999. F0-contours in emotional speech. In: Proceedings of the 14th International Conference of Phonetic Sciences. San Francisco, U.S.A., pp. 929–932.
Proctor, M., Bone, D., Katsamanis, N., Narayanan, S.S., 2010. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In: Proceedings of Interspeech. ISCA, pp. 1576–1579.
Ren, G., Zhang, X., Duan, S., 2018. Articulatory-acoustic analyses of Mandarin words in emotional context speech for smart campus. IEEE Access 6, 48418–48427.
Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. Trans. Acoust. Speech Signal Process. 26 (1), 43–49.
Santos, J., Wright, G., Pauly, J., 2004. Flexible real-time magnetic resonance imaging framework. In: Proceedings of Engineering in Medicine and Biology Society. IEEE, pp. 1048–1051.
Scherer, K.R., 1986. Vocal affect expression: a review and a model for future research.. Psychol. Bull. 99 (2), 143–165.
Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Commun. 40 (1−2), 227–256.
Smith, D.R.R., Patterson, R.D., 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J. Acoust. Soc. Am. 118 (5), 3177–3186. https://doi.org/10.1121/1.2047107.
Story, B.H., 2009. Vowel and consonant contributions to vocal tract shape. J. Acoust. Soc. Am. 126 (2), 825–836. https://doi.org/10.1121/1.3158816.
Story, B.H., Titze, I.R., 1998. Parameterization of vocal tract area functions by empirical orthogonal modes. J. Phon. 26 (3), 223–260. https://doi.org/10.1006/jpho.1998.0076.
Story, B.H., Titze, I.R., Hoffman, E.A., 1996. Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am. 100 (1), 537–554.
Tartter, V.C., 1980. Happy talk: perceptual and acoustic effects of smiling on speech. Percept. Psychophys. 27 (1), 24–27. https://doi.org/10.3758/BF03199901.
Xu, Y., Chuenwattanapranithi, S., 2007. Perceiving anger and joy in speech through the size code. In: Proceedings of the International Conference on Phonetic Sciences, pp. 2105–2108.
Yehia, H.C., Takeda, K., Itakura, F., 1996. An acoustically oriented vocal-tract model. IEICE Trans. Inf. Syst. E79-D (8), 1198–1208.